Performance of a Disk Storage System at a Tier-2 Site

International Center for Elementary Particle Physics (ICEPP),
the University of Tokyo

Hiroyuki Matsunaga

ISGC2009
Taipei, Taiwan
April 2009

Outline

- Tokyo regional center
- Disk storage system
- Data transfer over WAN
  - gridFTP
- Data access in LAN
  - Posix-like protocols (rfio, xrootd), nfs
- Summary

Tokyo Regional Center

- Located at ICEPP, the University of Tokyo
- Supports only ATLAS experiment
- Resources are separated into Grid and Non-Grid portions
  - Grid: WLCG Tier-2 site (TOKYO-LCG2)
    - 120 Worker Nodes (480 cores)
    - ~400 TB disk storage (13 disk servers)
      - DPM (Disk Pool Manager) as Storage Management system
  - Non-Grid: Local resource for Japanese users
    - ~500 batch nodes (~2000 cores)
    - 200~300 TB disk storage
      - No Storage management system at present. Just using NFS (and FTP).
      - Will deploy HSS with tape system

Tokyo Tier-2 site

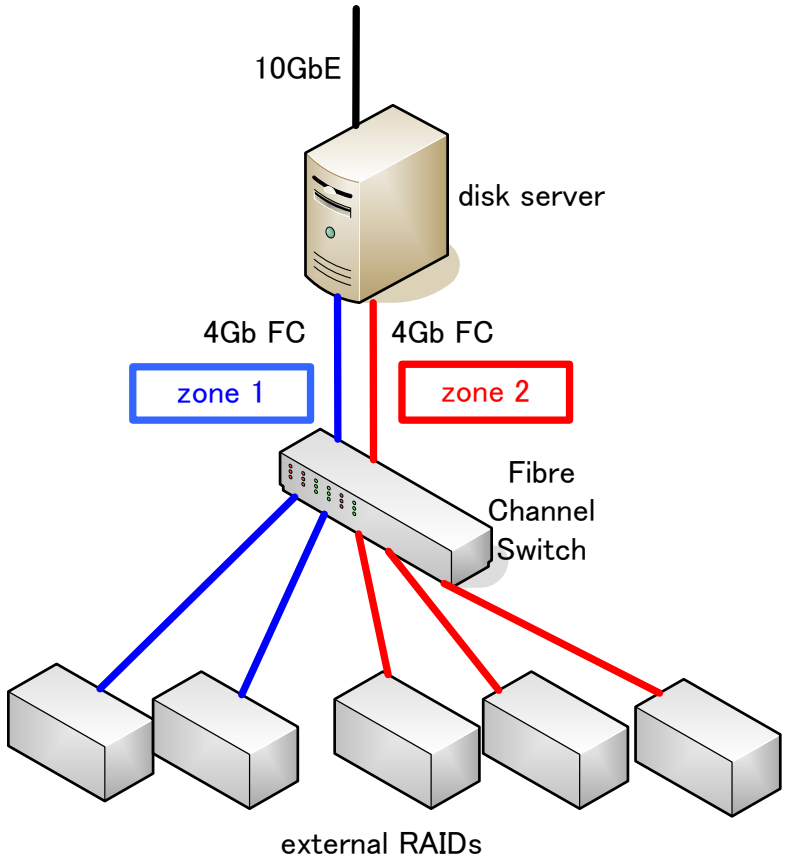
- gLite middleware (developed by EGEE) deployed
  - Operation supported by Asia-Pacific ROC
- Unique ATLAS (Tier-2) site in Japan
- Roles of Tier-2 site in ATLAS
  - MC data production
    - Input data from the Tier-1
    - Output data to the Tier-1
  - User analysis
    - Input data from the Tier-1
- CC-IN2P3 (Lyon, France) is the Tier-1 for Tokyo site
  - MC data production and data transfer are coordinated by the French team

DPM architecture

- Servers (on head node)
  - SRM, DPM, DPNS (xrootd: option)
  - MySQL database backend
- Storage services (on disk server)
  - gridFTP
  - rfiod
  - (xrootd: option)
- Our configuration is 1 head node and 13 disk servers

Hardware setup

- Head node
  - Blade server
    - 2 x dual-core CPU (Xeon 5160)
    - 16GB RAM
    - 73GB local disk (mirror)
    - 1GbE
- Disk server
  - 2U rackmount server
    - 2 x dual-core CPU (Xeon 5160)
    - 8GB RAM
    - 73GB local disk (mirror)
    - 10GbE
    - 2-port 4Gb Fibre Channel HBA
- Fibre Channel Switch (4Gb)
- External RAIDs
  - 4Gb Fibre Channel IF
  - 16 x 500 GB SATA HDD
  - RAID-6 (~7TB/box)
- 5 RAIDs are connected to a disk server



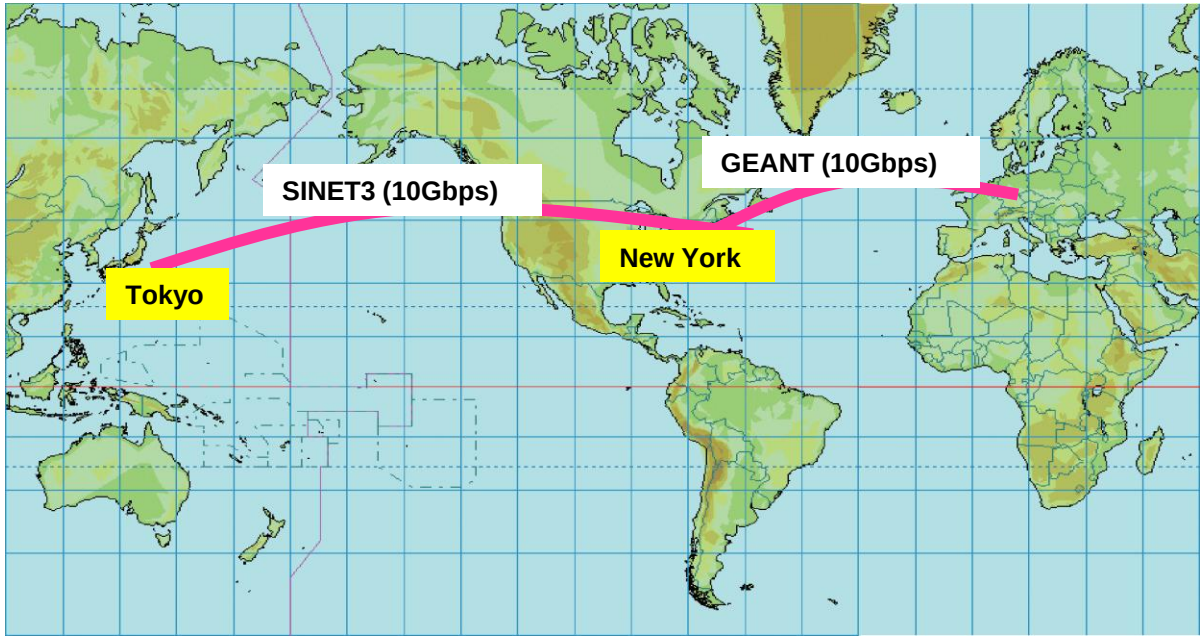
Software

- OS: Scientific Linux CERN (SLC) 4
  - kernel 2.6.9, 64bit
  - filesystem: XFS
    - 1 filesystem per RAID (~7TB/filesystem)
  - 10GbE driver (Chelsio cxgb) installed by hand
- DPM 1.6.x (in gLite 3)
  - All 65 (13 x 5) filesystems (~400TB) in 1 pool
    - When writing a file, a filesystem is selected with simple round robin
      - Files spread equally over the filesystems
- System tuning has not been done seriously
  - Some settings (TCP window size etc.) done by YAIM
- Server Performance
  - Network: Up to ~750MB/s between two disk servers
  - Disk: 300MB/s single sequential write, 180MB/s read

Data transfer over WAN

Data transfer

- Main data transfer channel is from Tier-1 to Tier-2
- Due to large round trip time (290ms between Lyon and Tokyo), fast data transfer is difficult to achieve
  - 10Gbps available via academic links (shared with other traffic)

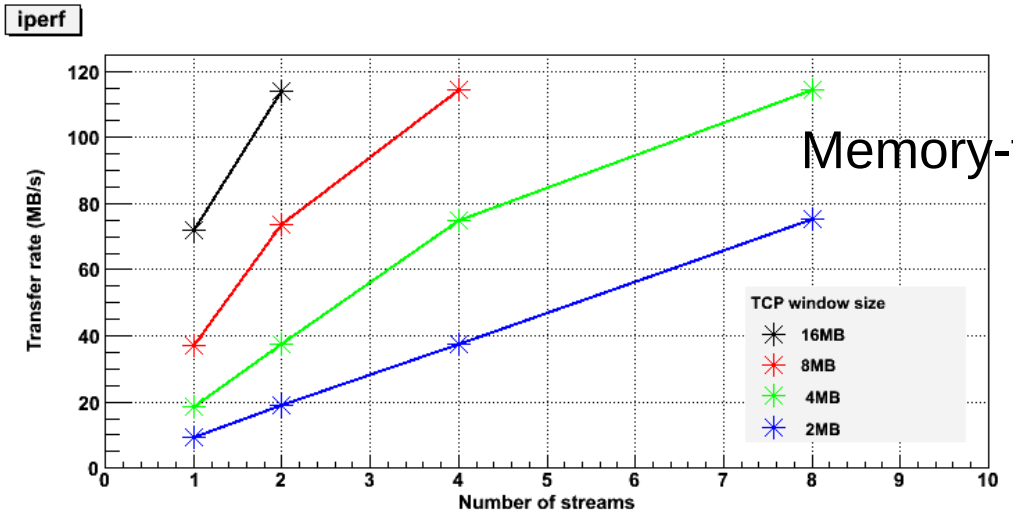


Test of data transfer

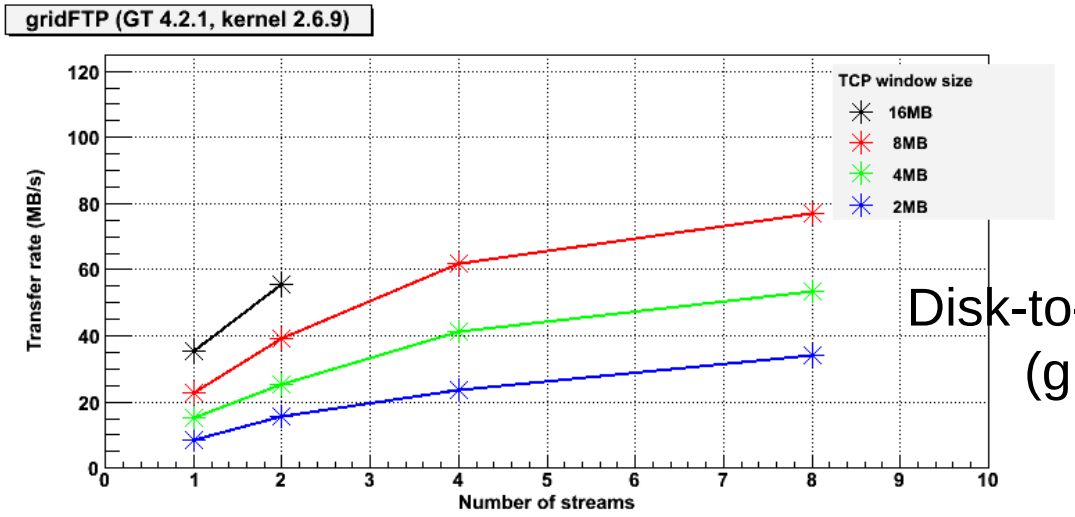
- Test node set up at CERN and Tokyo
  - Almost same path as Lyon-Tokyo
  - Bandwidth limit at CERN router and NIC of test node at CERN (1Gbps)
  - Use a spare disk server at Tokyo
- Important parameters for the fast data transfer
  - TCP window size (2MB at DPM production site)
  - Number of streams per file transfer (10)
  - Number of concurrent files (20 for 13 disk servers)
- File size is >1GB in this test

Test Results (1 file transfer)

- Varied TCP window size and number of streams
- As expected, gridFTP is slower than iperf
  - 70MB/s vs. 35MB/s in case of 2MB and 8 streams



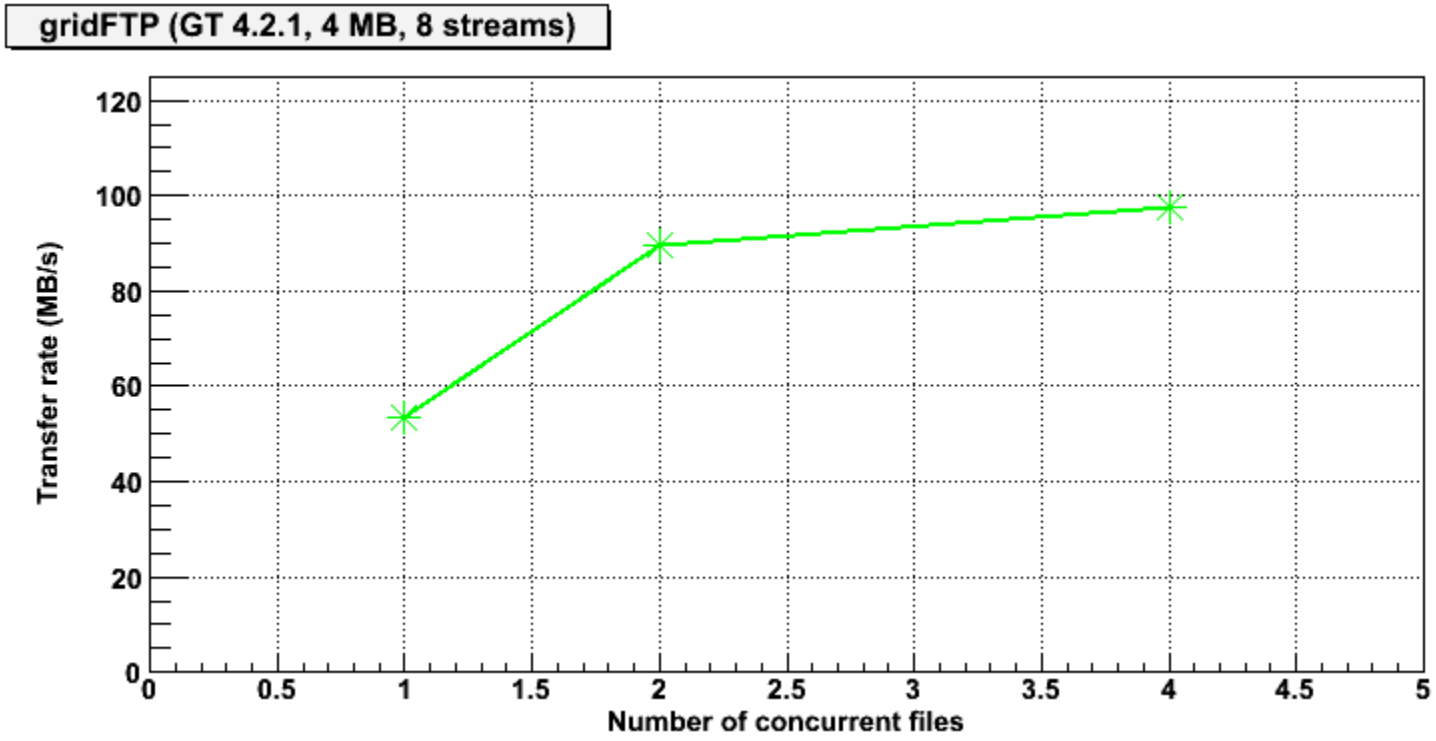
Memory-to-Memory (iperf)



Disk-to-Disk (gridFTP)

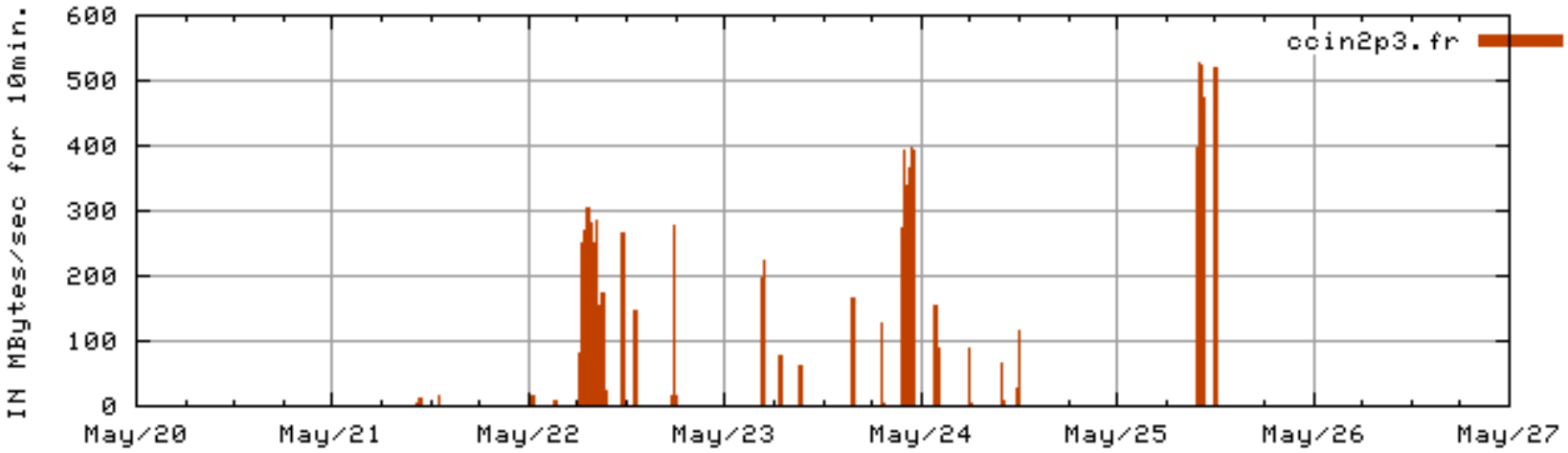
Multiple file transfer

- All files reside in a filesystem on both sender and receiver nodes



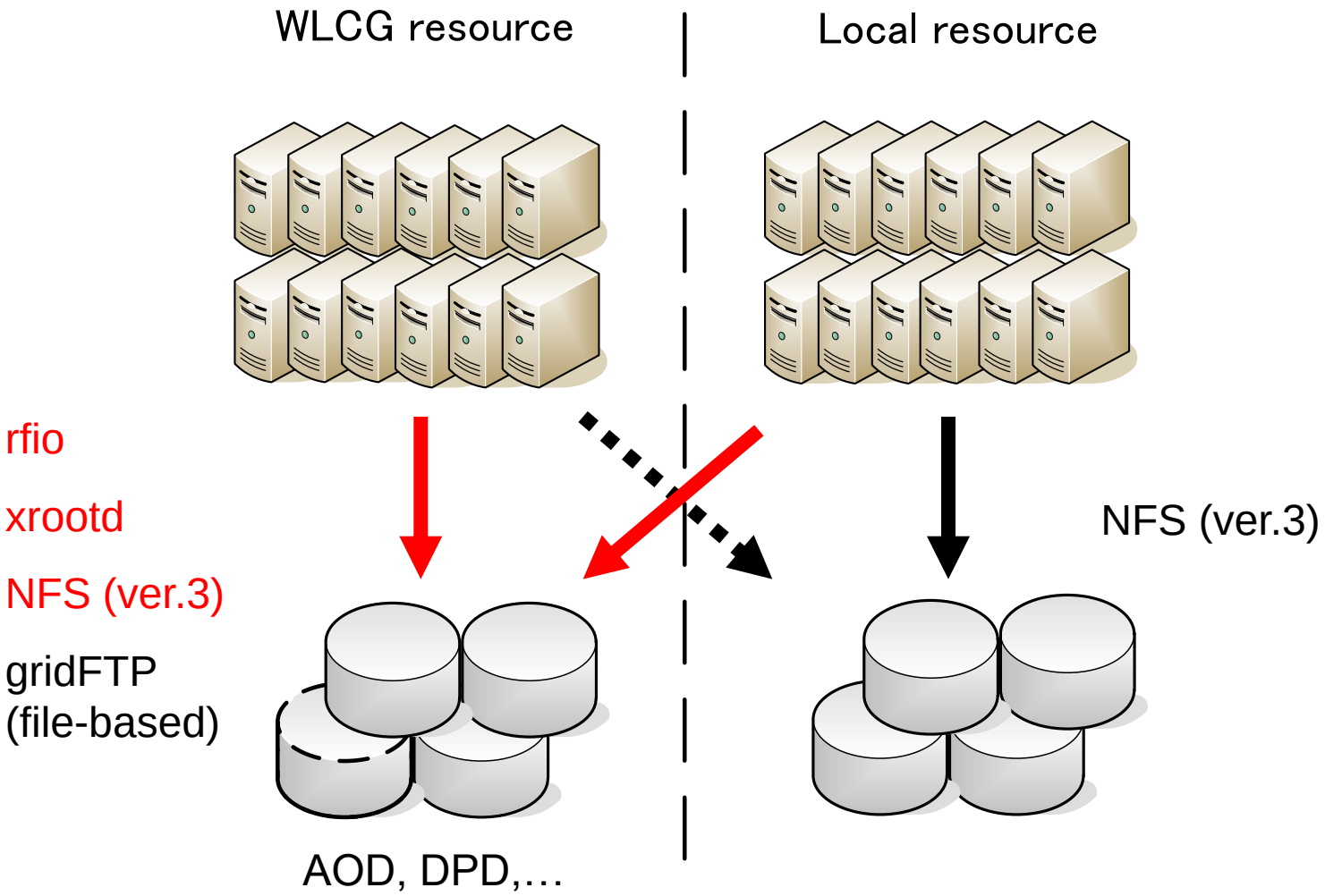
Data transfer from Lyon to Tokyo using production system

- >500MB/s is the best record so far
  - During CCRC08 (3.5GB/file)
  - Number of disk servers: >30 at Lyon, 6 at Tokyo
  - Low activities for other jobs at Lyon
- 300~400 MB/s observed from time to time



Data access in LAN

Data access in LAN



Comparisons

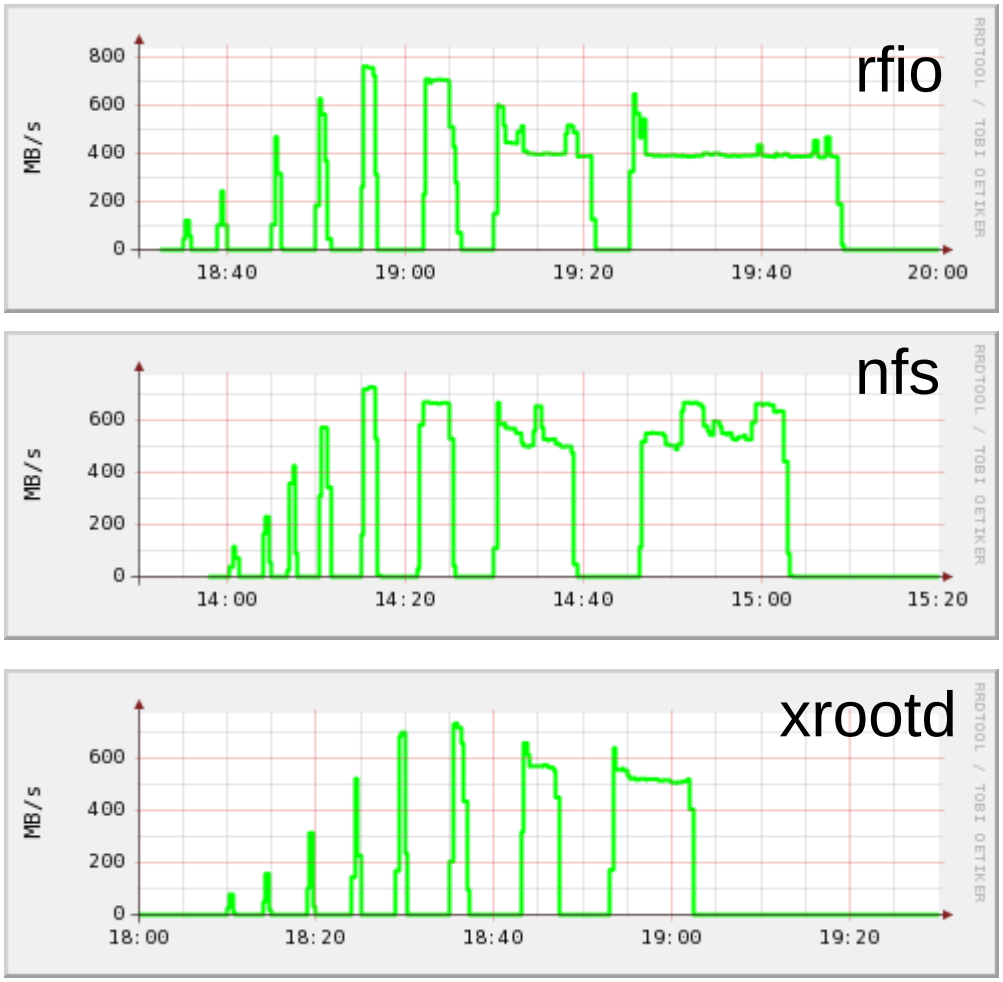
- rfio
  - Has been used at CERN for a long time
  - Included in DPM and CASTOR
- xrootd
  - Low latency, no third party software needed
  - Not installed in DPM by default
  - Still problem in GSI support ?
  - Cannot handle a file >2GB ?
- NFS (ver.3)
  - No name service
  - Problem in user mapping and security
  - Only for local users ?
- gridFTP
  - Retrieval of a whole file before processing
  - Need scratch space on worker node
- In ATLAS, gridFTP and rfio are used for user analysis tests

Test Setup

- Installed DPM and DPM-xrootd on a test node (single disk server)
  - Same configuration as the production disk server
    - Attached 5 RAIDs / filesystems
  - All services running on this node
  - Running 64 nfsd services + xrootd
- Each client copies a file from the test server to the local disk on WN
  - Using rfcp (rfio), xrdcp (xrootd), or cp (nfs)
  - 1 client on 1 WN (1GbE)
  - Each client copies a different file
  - Clients simultaneously start copying files
    - 4GB file for rfio and nfs. 2GB for xrootd due to the file size limit.

Test Results

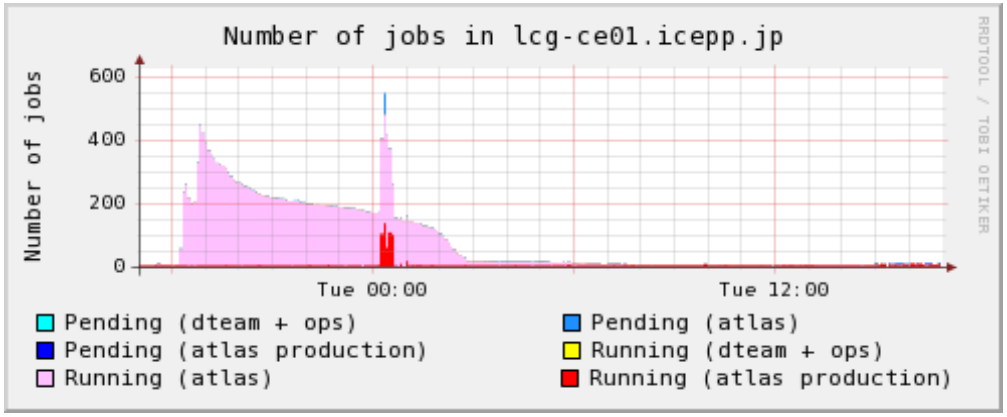
Total transfer rate

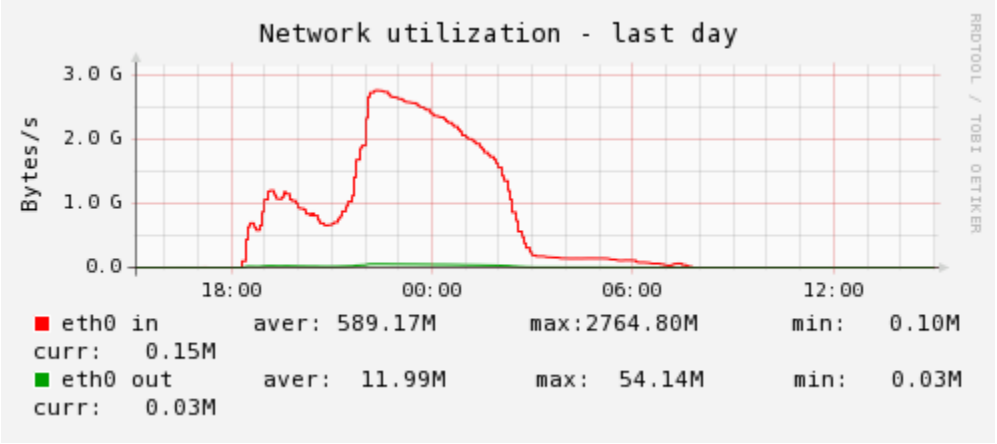


- 1, 2, 4, 8, 16, 32, 64 and 128 clients
- 1 client failed with timeout in rfio mode with 128 clients
- “Tx ring full” message in server’s syslog in xrootd mode with 32 clients
- Peak rate of 700~750 MB/s observed with 16 or 32 clients
  - Limit of 10GbE
- Worse performance in rfio mode with many clients (saturated at 400MB/s)

ATLAS stress test

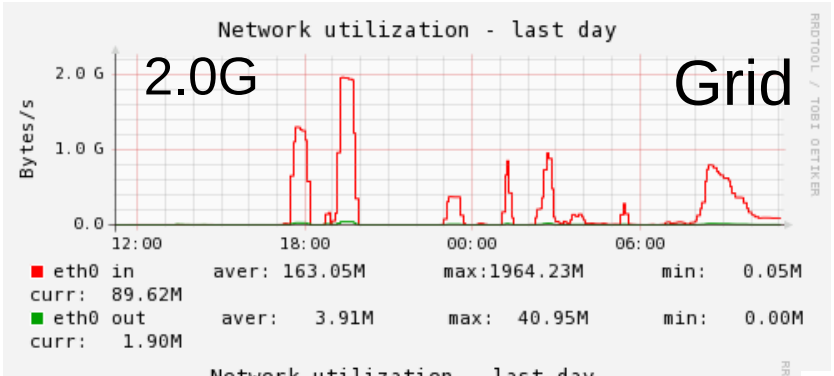
- “Hammercloud” test
  - Submit user analysis jobs into a “cloud” (Tier-1 + associated Tier-2 sites)
  - Physics analysis jobs reading many input data files
- Heavy load of name resolution (logical to physical address conversion) at the job start-up
- ~2.8GB/s of peak transfer rate with 13 disk severs
- Best performance among the ATLAS Tier-2 sites
  - CPU time / wall time
  - Event processing rate

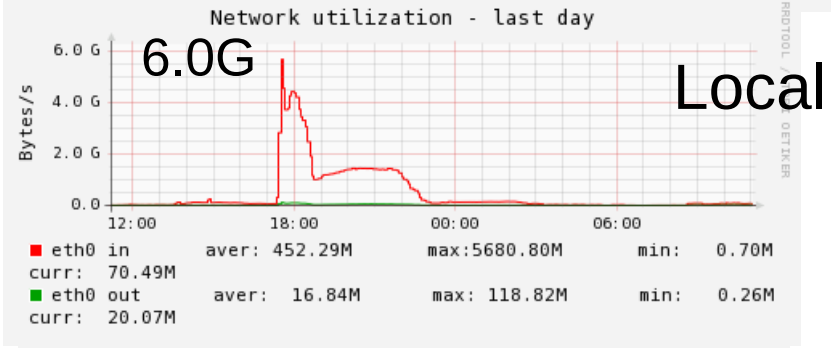


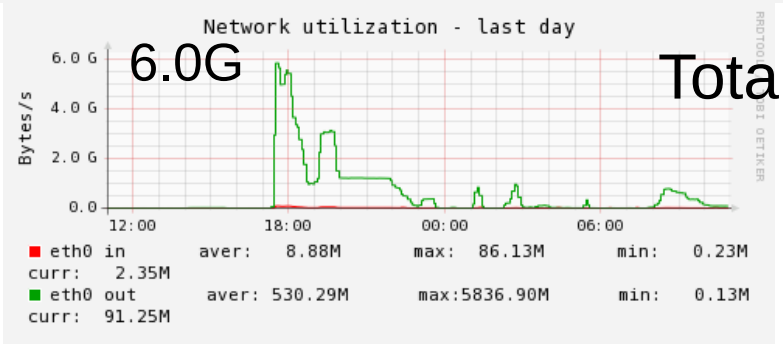


Real-life experience

- Both Grid and local worker nodes tried to access DPM data with rfio at the same time
- Total transfer rate up to 6GB/s
  - ~500MB/s per disk server
  - Saturated an interconnection between Ethernet switches







Summary

- Tokyo regional center
  - Grid and local resources for ATLAS
- Performance of the disk storage system was tested with a test server, and measured in production system
  - Almost sufficient for both WAN data transfer and LAN data access even without system optimization
  - There seems still room for improvement
  - Need more bandwidth between the Ethernet switches
  - Concern about performance of the backend MySQL
    - Will study in the future
- Next procurement in process
  - Faster CPU, RAID controller, …
  - Larger HDD (500GB -> 1 or 1.5TB)
  - 16 -> 24? HDDs/RAID
  - Will reduce number of RAIDs per disk server (5 -> 2?)